认知诊断缺失数据处理方法的比较：零替换、多重插补与极大似然估计法*

宋枝璘[1] 郭磊[1,2] 郑天鹏[3]

([1] 西南大学心理学部，重庆 400715) ([2] 中国基础教育质量监测协同创新中心西南大学分中心，重庆 400715)

([3] 北京师范大学中国基础教育质量监测协同创新中心，北京 100088)

摘 要 数据缺失在测验中经常发生，认知诊断评估也不例外，数据缺失会导致诊断结果的偏差。首先，通过模拟研究在多种实验条件下比较了常用的缺失数据处理方法。结果表明：①缺失数据导致估计精确性下降，随着人数与题目数量减少、缺失率增大、题目质量降低，所有方法的 PCCR 均下降，Bias 绝对值和 RMSE 均上升。②估计题目参数时，EM 法表现最好，其次是 MI，FIML 和 ZR 法表现不稳定。③估计被试知识状态时，EM 和 FIML 表现最好，MI 和 ZR 表现不稳定。其次，在 PISA2015 实证数据中进一步探索了不同方法的表现。综合模拟和实证研究结果，推荐选用 EM 或 FIML 法进行缺失数据处理。

关键词 认知诊断，GDINA 模型，缺失数据，多重插补，极大似然估计

1 引言

认知诊断评估(Cognitive Diagnosis Assessment, CDA)是最新一代的心理与教育测评技术，可以对个体认知过程、加工技能或知识结构进行诊断与评估。在 CDA 实施过程中，无法避免出现数据的缺失。已有研究表明，随着数据缺失率的增大，题目参数的估计精度及被试知识状态(Knowledge State, KS)的判准率均会下降，而选用不同的缺失值处理方法也会对模型拟合与参数估计带来不同影响(Dai, 2017; Pan & Zhan, 2020)。因此，在实际 CDA 测验中需要重视缺失数据问题，并选用合适方法处理，以提升诊断精度及题目参数估计精度。

目前，缺失数据的处理方法主要包括两大类：一是传统处理方法，如具有代表性的零替换(Zero Replace, ZR)方法。ZR 法操作便捷，在处理大规模数据时非常快速，在绝大多数统计软件上均可实现，并且不会造成被试的大量流失。因此，ZR 是研究者经常选用的方法之一，在 CDA 中也有使用（Aryadoust & Goh, 2001; Lee et al., 2011），且 ZR 方法目前被较多大型教育评估，如 PISA、TIMSS、PIRLS 所采纳(Xiao & Bulut, 2020)。虽然传统方法

收稿日期：2021-06-10

* 国家自然科学基金青年项目(31900793); 北京师范大学中国基础教育质量监测协同创新中心重大成果培育性项目(2019-06-023-BZPK01); 中央高校基本科研业务费专项资金(SWU2109222)资助。

通讯作者：郭磊, E-mail: happygl1229@swu.edu.cn

比较便捷，但会导致统计效力和参数估计精度的下降，因此有研究者并不建议使用(Dong & Peng, 2013; Enders, 2010)。第二类是基于模型的处理方法，近年来，随着统计技术不断发展，基于模型的处理方法相继被提出，并被证明其处理效果优于传统方法，因此这些方法越来越受到重视。其中，极大似然估计(Maximum Likelihood Estimation, MLE)和MI(Multiple Imputation, MI)方法的应用最广泛(Rotnitzky, 2008; Schafer & Graham, 2002)。MLE 是通过加工似然函数对缺失数据进行处理，包括期望最大化算法(Expectation-Maximization algorithm, EM)和全息极大似然估计方法(Full Information Maximum Likelihood, FIML)。对于 FIML、EM 和 MI 三种方法，均有研究证明其表现优于传统方法(Graham, 2009; Jeličić et al., 2010;Van Buuren, 2018 ;Wothke, 2000)。本文所采用方法的具体介绍请参见 2.2 部分。

CDA 中探讨缺失值及其处理的研究中，一部分研究者仅探讨了缺失数据对诊断结果的影响，如 Xu 和 von Davier(2006)的研究表明：当数据缺失率达 50%时，认知诊断模型仍能得到较好的估计结果。但该研究未考虑不同的缺失机制，且仅考虑了缺失数据对参数估计的影响而未考虑缺失数据处理方法本身对结果产生的影响。Pan 和 Zhan(2020)在纵向诊断中探讨了缺失数据对诊断精度的影响，也得到相似的研究结果。Dai(2017)首次探讨了不同的缺失值处理方法在 CDA 中的表现，作者在 DINA(Deterministic Inputs, Noisy “and” Gate)模型(Junker & Sijtsma, 2001)基础上，比较 EM 和一些传统方法的表现，结果表明：在估计被试 KS 时，EM 在多数条件下表现较好；在题目参数估计时，EM 和传统方法的表现随条件改变而各不相同。

尽管已有上述文献研究了 CDA 中的缺失数据问题，但过往研究首先未曾考虑在缺失数据分析领域中表现较好、应用广泛的 MI 和 FIML 方法。其中，MI 法已被证明其表现较为优异和稳健(Van Buuren, 2018; Schafer & Graham, 2002)，且于近年来被广泛用于缺失数据的处理中(Leacy et al., 2017; Rezvan et al., 2015)。FIML 采用“一步式”操作，直接使用带缺失值的作答数据进行模型拟合，比其它基于模型的方法更便捷(Graham, 2009; Schafer & Graham, 2002)，此外，基于模型的方法表现更加出色，但在不同研究背景下的表现有较大差异，取决于具体的模型、数据和条件(Newman, 2003; Dai, 2017)。因此，有必要在 CDA 中系统地探索这些基于模型方法的表现，并与传统方法进行比较。

基于系统全面比较缺失值处理方法这一主旨，本研究还做了如下推进：①Dai(2017)采用的 DINA 属于简约模型，它的非补偿模型特点往往与现实测验情景不符。而饱和模型，如 GDINA(Generalized Deterministic Inputs, Noisy “and” Gate)模型(de la Torre, 2011)等受到了

较多关注，并应用于多数研究中(Bai, 2020; 高旭亮等, 2018)，GDINA 不仅包含属性主效应，还将属性间交互作用考虑在内，更加符合现实情况，对实际测验拟合更佳，对 GDINA 及 DINA 模型的介绍及含义参见 2.1 部分。②现有诊断测验中比较缺失数据处理方法的研究仅使用了模拟研究，但模拟研究的生态效度如何并未在实证数据中得到检验，因此结果是否能进一步推至实际情况有待进一步验证。③除了 MI 和 FIML，本文还选取了传统方法中具有代表性的 ZR 法，以及插补后可以得到无偏估计结果且在处理 CDA 及其它测验类型的数据缺失值时，表现较为优异的 EM 方法(Dai. 2017; Newman, 2003; Rasmussen, 2007)。

综上，本研究的主要目的是将 MI 和 MLE 法引入 CDA 中，对不同缺失数据处理方法进行全面比较，并提出实践中处理缺失数据的建议。下文首先对认知诊断模型和各缺失数据处理方法进行简单介绍。其次，通过模拟研究，在不同实验条件下探究了各缺失数据处理方法的表现。第三，以 PISA2015 年基于计算机测评中的数学素养为例，比较不同缺失数据处理方法在实证数据中的效果，验证不同方法的生态效度。最后，我们讨论了研究结果及未来研究的发展方向。

2 认知诊断模型及缺失数据介绍

2.1 认知诊断模型

本研究所采用的诊断模型为 GDINA，其表达形式见公式(1):

$$P\left(Y_{ij}=1\middle|\boldsymbol{\alpha}_{ij}^{*}\right)=\delta_{j0}+\sum_{k=1}^{K_j^*}\delta_{jk}\alpha_{ik}+\sum_{k'=k+1}^{K_j^*}\sum_{k=1}^{K_j^*-1}\delta_{jkk'}\alpha_{ik}\,\alpha_{ik'}+\dots+\delta_{j1,\dots,K_j^*}\prod_{k=1}^{K_j^*}\alpha_{ik} \quad (1)$$

在 GDINA 模型中，被试在每道题目上被归为$2^{K_j^*}$个潜类别，其中$K_j^*=\sum_{k=1}^{K}q_{jk}$，表示题目 j 所考察的属性数量，q_{jk}=1 表示题目 j 考察了属性 k。$\boldsymbol{\alpha}_{\rm ij}^{*}=\left(\alpha_{ij1},\dots,\alpha_{ijK_j^*}\right)$为在被试属性向量$\boldsymbol{\alpha_{ij}}=\left(\alpha_{ij1},\dots,\alpha_{ijK_j}\right)$基础上，仅保留题目 j 所考察属性，形成的坍塌(collapse)属性向量(K_j为测验考察的所有属性个数)。δ_{jo}为题目 j 的截距项，即当被试未掌握题目所考察属性时正确作答的基线参数。δ_{jk}为属性 k 的主效应，表示当被试仅掌握某一属性 k 时，对正确作答概率的影响。$\delta_{jkk'}$是题目 j 在属性 k 和 k'上的二阶交互效应，表示同时掌握两个属性对正确作答概率的影响。$\delta_{j1,\dots,K_j^*}$为题目 j 在属性 $1, 2, \dots, K_j^*$上的最高阶交互作用，表示掌握了题目 j 考察的所有属性时，对正确作答概率的影响。其中，截距项δ_{jo}衡为非负数，主效应项为非负数，而交互作用项可以取任意值。

GDINA 模型属于饱和模型，对 GDINA 进行约束，即仅保留公式（1）中的截距项和

最高阶交互项，便可得到 DINA 模型：$P\left(Y_{ij}=1\middle|\boldsymbol{\alpha}_{ij}^{*}\right)=\delta_{j0}+\delta_{j1,\dots,K_j^*}\prod_{k=1}^{K_j^*}\alpha_{ik}$。其含义为：当且仅当被试 i 掌握了题目 j 考核的所有属性时，该被试倾向于答对这道题目；而当被试 i 未掌握题目 j 考核的所有属性时，即认为该被试倾向于答错这道题目。

2.2 缺失数据机制介绍

缺失数据可以通过缺失机制进行分类，Rubin(1976)定义了三种缺失的数据机制；完全随机缺失(missing completely at random, MCAR)，随机缺失(missing at random, MAR)和非随机缺失(missing not at random, MNAR)。在 MCAR 机制下，数据的缺失是完全随机的，不依赖于任何变量，即不论其它变量（如题目难度、区分度、被试能力值等）如何变化，数据产生缺失的概率都是均等的；在 MAR 机制下，数据缺失的概率并不是随机的，会受到数据集中已观测到的、不含缺失值的变量（如被试年龄、能力值等）的影响，但不受缺失数据自身的影响；在 MNAR 机制下，数据缺失的概率与缺失变量本身相关，如某一问题设计的过于敏感造成的缺失。

在心理教育测评中，这三种缺失数据的机制都有可能存在。Huisman 和 Molenaar(2001)认为，测评中缺失的作答是由学生无意中报告的，因此将测评中的缺失数据视为 MCAR 机制下的缺失；还有研究者假设测评中存在 MAR 机制，因为数据的缺失与特定的个体特征有关(De Ayala el al., 2001; Finch, 2008) ;还有研究表明在某道题目上数据的缺失是受到题目本身特征的影响，即存在 MNAR 缺失机制(Brown et al., 2014)。

2.3 缺失数据的处理方法

依据前文综述，本研究选取了常见且被广泛使用的传统方法 ZR 法(Aryadoust & Goh, 2001; Lee et al., 2011)。基于模型的缺失数据处理方法中应用最为广泛(Rotnitzky, 2008; Schafer & Graham, 2002; Leacy et al., 2017; Rezvan et al., 2015)，处理缺失值效果更具优势(Rasmussen, 2007; Graham, 2009; Jeličić et al., 2010; Wothke, 2000)并且适用于二分变量插补(Marshall et al., 2010; Van Buuren, 2018)的 MI-PMM、MI-CART、MI-LOGREG.BOOT、EM 和 FIML 这几种方法。

2.3.1 零替换(ZR)

零替换，即将缺失的作答视为错误回答，用“0”值替换缺失数据。再将替换好的完整数据集输入模型，进行分析。

2.3.2 多重插补(MI)

多重插补(Rubin, 1976)是一种基于重复模拟的缺失数据处理方法，MI 包括三个步骤：

插补(imputation)、分析(analysis)和合并(pooling)。首先，MI 依据具体的插补模型(MI-PMM、MI-CART、MI-LOGREG.BOOT)对缺失数据进行多次插补，最终得到多个经插补后的完整数据（最好是 20 个或更多；Graham et al., 2007）。然后依照模型（如线性模型、广义线性模型等）对这 20 或更多个完整数据集进行分析，依据 Rubin 规则计算各完整数据的参数估计值，最后将参数估计最佳的插补结果输出（Mazza et al., 2015）。最终输出的插补结果将作为完整数据集输入模型，进行分析。

由于本研究考虑的是二级计分形式，且具备局部独立性，因此在使用 MI 对作答矩阵进行插补时，分别选择各插补模型对数据进行多次插补（m=20），并从插补好的数据集中随机抽取一个完整数据集作为插补结果。MI 系列方法中的分类回归树方法(Classification and regression trees, MI-CART)、预测均值匹配(Predictive mean matching, MI-PMM)和自助比率对数回归(Logistic regression with bootstrap, MI-LOGREG.BOOT)均适用于二分变量插补，且在处理缺失数据时表现较好(Marshall et al., 2010; Van Buuren, 2018)，因此本研究主要选择了这三种 MI 模型。MI 系列方法的具体公式和详细操作步骤可参见 Van Buuren(2018)书中的具体介绍，下面对各方法原理与基本步骤进行介绍。

（1）MI-PMM：PMM 即预测均值匹配，它根据指定的回归模型计算缺失值的预测值，从而进行插补。已知 Y 为作答矩阵，记 Y'为删除 Y 中作答存在缺失值的被试后，由所有不含缺失值被试的作答数据构成的矩阵。PMM 大致步骤如下：1) 使用 Y'数据，建立多元回归模型，估计得到回归参数。2) 对第一步中得到的回归参数进行修正，得到一个适用于 Y 中所有被试的回归模型，并使用这一回归模型计算出所有被试的估计值。3) 针对每一个存在缺失数据的被试，匹配多个估计值与其估计值近似且不含缺失数据的被试，构成捐赠者库，从捐赠者库中随机抽取数值替换缺失数据，实现对缺失数据的插补。此方法假设缺失值的分布与候选数据集相同，并使用已有数据中非缺失部分对缺失值进行插补，从而避免了无意义插补的问题（例如，身高为负值）(Van Buuren, 2018)。

（2）MI-CART：CART 即分类回归树，是一种回归与分类技术。该方法一般使用递归划分法构建预测模型，将待处理的所有变量划分为尽可能同质的类别，最终形成决策树，并从与缺失作答相似的节点中随机抽取完整作答，对缺失数据进行插补。其操作步骤如下：1）使用递归法对作答矩阵中的数据进行多次切分，模型将选择具有最佳分裂和最均匀子群的切分点作为最佳切分点，将数据切分为两个子节点。2）多次重复第一步，直至数据不可再分。此时，每一个子节点下的数据均同质，经切分得到的结果即为作答矩阵的分类回归树。3）对每一个缺失作答，根据分类回归树找到它所属的终端节点，并从该节点中随机抽

取作答对缺失数据进行插补。

（3）MI-LOGREG.BOOT：该方法使用基于 Bootstrap 的贝叶斯逻辑回归模型对缺失数据进行插补。Bootstrap 法的原理是以样本代表总体，在样本中进行有放回抽样，每次重复抽取 *n* 个数据组成一个样本，重复这一过程多次得到多个样本，最后基于这些样本进行统计计算。MI-LOGREG.BOOT 通过贝叶斯逻辑回归模型进行，对经 bootstrap 法处理后的作答矩阵进行回归分析，通过计算和比较拟合所得参数，选取拟合结果最佳的数据对缺失值进行插补(Van Buuren, 2018)。

2.3.3 期望最大化算法(EM)

EM 算法是一种通过计算极大似然对缺失进行处理的迭代算法(Dempster et al., 1977)。此方法原理是认为存在一个估计参数，与缺失数据相关且可以互相推导。因此给定估计参数的初始值，即可以通过不断迭代对缺失数据进行插补。每一次迭代包括了 E 步和 M 步。E 步即期望步，依据现有数据和前一次迭代所得到的估计参数，对缺失数据进行填补，并计算其对数似然函数的条件期望；M 步即极大化步，用极大化对数似然函数进一步确定估计参数的值，并用于下一步迭代。算法在 E 步和 M 步之间不断迭代，直至两次迭代之间的参数变化较小时结束(叶素静等, 2014)。EM 的具体公式和详细操作步骤可以参考 Dempster 等(1977)研究中的具体介绍，下面对其原理与基本步骤进行介绍。定义Y_{obs}为已观测到的、未缺失的数据，定义Y_{mis}为缺失数据，则含有缺失数据的作答矩阵 Y=（Y_{obs}, Y_{mis}），EM 法就是使用待估参数θ和Y_{obs}，对Y_{mis}进行插补。EM 方法在 CDA 中的实现过程为：

1) 首先给定参数θ的初始值，θ是一系列用于定义 Y 分布的参数合集。例如通过定义均值和方差，假设作答矩阵 Y 为正态分布，则此时初始估计参数θ可记作：$\theta=(\mu, \sigma^2)$。

2) E 步：对于Y_{mis}中的任一缺失数据y_{ij}，使用第一步给定的θ和Y_{obs}，计算Y_{obs}的条件期望值，并用这一期望值代替缺失数据，如公式(2)所示。其中 *i* 为被试，*j* 为题目，*t* 为迭代次数。

$$y_{ij}^{(t)} = E(y_{ij}|Y_{obs}, \theta^{(t)}) \qquad (2)$$

3) M 步：使用似然函数 Q 与经 E 步处理后的作答矩阵 Y，计算最大似然的参数估计值，如公式(3)所示，并选取满足最大似然值的参数值θ，作为新一轮迭代中的参数估计值。

$$Q(\theta^{(t+1)}|\theta^{(t)}, Y) = \max\left(Q\left(\theta\middle|\theta^{(t)}, Y\right)\right) \qquad (3)$$

4) 不断重复步骤 2-3 直到参数估计结果收敛，例如前后两次迭代之间参数估计的变化量：$\|Q(\theta^{(t+1)}|\theta^{(t)}, Y) - Q(\theta^{(t)}|\theta^{(t-1)}, Y)\|$小于特定值（如 .0001）。

同 MI 法相同，将 EM 处理后的完整数据集输入模型，进行后续的分析。

2.4.4 全息极大似然估计算法(FIML)

与删除方法不同，FIML 不排除缺失作答的情况。包括不完整案例的观察分数可以提高准确性，因为不完整变量与其他（完整或不完整）变量之间的关联会告知估计程序哪些参数值最有可能（Mazza et al., 2015）。FIML 使用缺失数据中所有的可用数据建立模型，并运用似然函数估计参数，对缺失数据进行处理(Eekhout et al., 2015)。例如，在运用 FIML 方法对包含缺失值的数据进行题目参数估计时，其似然函数的计算只连乘在该题目上有作答的数据的正确作答概率值，而未作答的数据不参与计算。不同于其它方法需要先插补再估计，FIML 仅需要一步，就可以同时实现缺失数据处理和参数估计过程，因此更加高效(Graham, 2009)。不同于以上几种方法，在本研究中，使用 R 中的 GDINA 包进行 FIML 方法的处理，FIML 方法为 GDINA 包的默认处理缺失值方法，即当输入的作答矩阵为包含缺失数据的矩阵时，软件默认使用 FIML 方法进行处理和模型的分析。

3 模拟研究

3.1 研究设计

为了充分探讨不同缺失数据处理方法在 CDA 中的表现，本研究采用 2✕3✕3✕3✕3✕6 的完全交叉实验设计，共包含六个自变量，其设置如下所示：

（1）被试数量：包括三个水平，200 人、400 人和 1000 人(Dai, 2017; de la Torre, 2011)；

（2）题目数量：包括两个水平：15 题和 30 题(Dai, 2017)；

（3）题目质量：参照 Ma 等人(2016)的设置包含三个水平：高质量、中等质量和低质量。题目为低质量时，参数设定为：$\mathrm{P}\left(Y_{ij}=1\middle|\boldsymbol{\alpha}_{\mathbf{ij}}^{*}=\mathbf{0}\right)\in U(0.05,0.15)$，$\mathrm{P}\left(Y_{ij}=1\middle|\boldsymbol{\alpha}_{\mathbf{ij}}^{*}=\mathbf{1}\right)\in U(0.85,0.95)$；题目为中等质量时，参数设定为：$\mathrm{P}\left(Y_{ij}=1\middle|\boldsymbol{\alpha}_{\mathbf{ij}}^{*}=\mathbf{0}\right)\in U(0.15,0.25)$，$\mathrm{P}\left(Y_{ij}=1\middle|\boldsymbol{\alpha}_{\mathbf{ij}}^{*}=\mathbf{1}\right)\in U(0.75,0.85)$；题目为高质量时，参数设定为：$P\left(Y_{ij}=1\middle|\boldsymbol{\alpha}_{ij}^{*}=\mathbf{0}\right)\in U(0.25,0.35)$，$P\left(Y_{ij}=1\middle|\boldsymbol{\alpha}_{ij}^{*}=\mathbf{1}\right)\in U(0.65,0.75)$。$P\left(Y_{ij}=1\middle|\boldsymbol{\alpha}_{ij}^{*}=\mathbf{0}\right)$表示被试 i 未掌握题目 j 考察的所有属性时，答对题目的概率。$\mathrm{P}\left(Y_{ij}=1\middle|\boldsymbol{\alpha}_{\mathbf{ij}}^{*}=\mathbf{1}\right)$表示被试 i 掌握了题目 j 考察的所有属性时，答对题目的概率。其中，Y_{ij}为被试 i 在题目 j 上的作答情况，$\boldsymbol{\alpha}_{ij}^{*}$为在被试原属性向量基础上，仅保留题目 j 所考察属性形成的坍塌属性向量。

（4）数据缺失机制：包括三种缺失机制：MCAR、MAR 和 MNAR(De Ayala, 2001; Finch, 2008)；

（5）数据缺失率：包括三个水平：10%、20%、30%(Dai, 2017)；

（6）缺失数据处理方法：ZR、MI-CART、MI-PMM、MI-LOGREG.BOOT、EM 和 FIML 方法；

其中，$\boldsymbol{Q}$ 矩阵设定参见附录一。

3.2 模拟过程

（1）完整数据生成：被试 KS 真值从多元正态分布中生成，即$\alpha \sim MVN(0_K, \Sigma)$，协方差设定为 0.5(如下所示)。被试作答数据使用 R 软件中 GDINA 包的 simGDINA()函数生成(Ma & de la Torre, 2020)。

$$\Sigma = \begin{bmatrix} 1 & \dots & .5 \\ \dots & \dots & \dots \\ .5 & \dots & 1 \end{bmatrix}$$

（2）缺失数据生成：缺失数据的生成参考 De Ayala 等(2001)和 Finch(2008)提出的方法，具体的生成过程参见附录二。

（3）缺失数据处理：使用 R 软件与 SPSS26.0 实现。首先，ZR 法通过自编 R 代码实现。MI 方法使用 R 软件中的 MICE 包(van Buuren & Groothuis-Oudshoorn, 2011)完成。为了保证处理效果，参照 Chen 等(2020)的研究将 MI 插补次数设定为 20 次。其次，在使用 EM 方法插补缺失数据时，我们首先采用了 R 软件中 TestDataImputation 包中的 EMimpute 函数，但在数据规模大、缺失比例较高（例：1000 人、30 题、30%缺失率）时，R 软件无法运行。因此，最终选用了 SPSS26.0 进行 EM 插补处理。FIML 方法通过 R 软件中的 GDINA 包完成。

3.3 评价指标

参数估计精度评价指标为偏差 Bias、均方根误差 RMSE(Ma & de la Torre, 2016)，计算见公式(2)和公式(3)。

$$Bias = \frac{\sum_{r=1}^{R}\sum_{c=1}^{2^K}\sum_{j=1}^{J}[\hat{P}^{(r)}(Y_J=1|\alpha_c)-P^{(r)}(Y_J=1|\alpha_c)]}{J\times 2^K\times R} \tag{4}$$

$$RMSE = \sqrt{\frac{\sum_{r=1}^{R}\sum_{c=1}^{2^K}\sum_{j=1}^{J}[\hat{P}^{(r)}(Y_J=1|\alpha_c)-P^{(r)}(Y_J=1|\alpha_c)]^2}{J\times 2^K\times R}} \tag{5}$$

其中，$\hat{P}^{(r)}(Y_J=1|\alpha_c)$ 表示 KS 为 α_c 的被试答对第 j 题的估计作答概率，$P^{(r)}(Y_J=1|\alpha_c)$ 表示 KS 为 α_c 的被试答对第 j 题的真实作答概率，R 表示总循环次数，r 表示当前循环次数。Bias 和 RMSE 越大，表明题目参数估计误差越大。

被试属性掌握的估计精度评价指标采用模式判准率(Pattern Correct Classification Rate,

PCCR)，计算见公式(4)。

$$PCCR = \frac{\sum_{r=1}^{R}\sum_{i=1}^{I} pm_{ir}}{R \times I} \qquad (6)$$

其中，I 为被试总数，pm_{ir}=1 表示第 r 次循环中被试的 KS 判断正确，反之表示判断错误。

3.4 模拟研究结果和讨论

由于 MAR 与 MCAR 机制下的结果基本相同，遂将 MCAR 的结果呈现于附录三。

3.4.1 MAR 机制的结果和讨论

（1）题目参数估计结果与讨论

图 1 和图 2 呈现了 MAR 机制下题目参数的估计结果，共包含 324 个条件。随被试数量和题目数量的增多，题目质量的提高和缺失率的降低，题目参数估计精度在提升。

整体来看，题目参数的估计偏差均较低，各方法表现均较好。其中，EM 法的表现最好，其后依次为 MI、FIML 和 ZR 法。EM 倾向于产生无偏估计，其 Bias 值分布范围为－0.003~0.001，非常接近 0；RMSE 分布范围为 0.014~0.100，是所有处理方法中值最小的。MI 也倾向于产生无偏估计，但其表现略差于 EM，其 Bias 分布范围为－0.003~0.001，RMSE 分布范围为 0.013~0.101。FIML 倾向于高估题目参数，其 Bias 值分布范围为－0.0001~0.008；RMSE 分布范围为 0.013~0.113。ZR 倾向于低估题目参数，其 Bias 值分布范围为－0.019~－0.003，均小于 0；RMSE 分布范围为 0.025~0.105。首先，不难看出，EM 和 MI 处理缺失数据后估计得到的题目参数精度最高，但其余方法的表现也不差，仅是相对而言略差，因为其中最大的 Bias 绝对值及 RMSE 值仅为 0.019 和 0.113。这一结果表明，这些方法处理缺失数据均能够得到较为理想的题目参数估计精度。其次，MI 系列中的三种方法的 Bias 和 RMSE 相似性较高，因此若要选用 MI 方法，MI 系列中的任一方法均可。最后，被试数量越多，题目质量越高，基于模型的方法表现越好，而 ZR 的表现则相反。这表明与传统插补方法相比，基于模型的方法更适合用于规模较大的测验情景。出现该结果的原因可能是基于模型的方法使用了数据中未缺失的作答信息进行建模，数据质量越好，规模越大，就越能从已有的作答数据中获得有效信息，从而提升了处理后的数据质量，使得估计结果更好。而 ZR 法将未作答信息全部用零值替换，这种替换会增大作答数据集当中的噪音，扭曲真实数据结构，当数据集越大，其噪音也就越大，最终导致估计精度下降。因此，在测验尤其是大规模测验中，应该使用基于模型的插补方法。

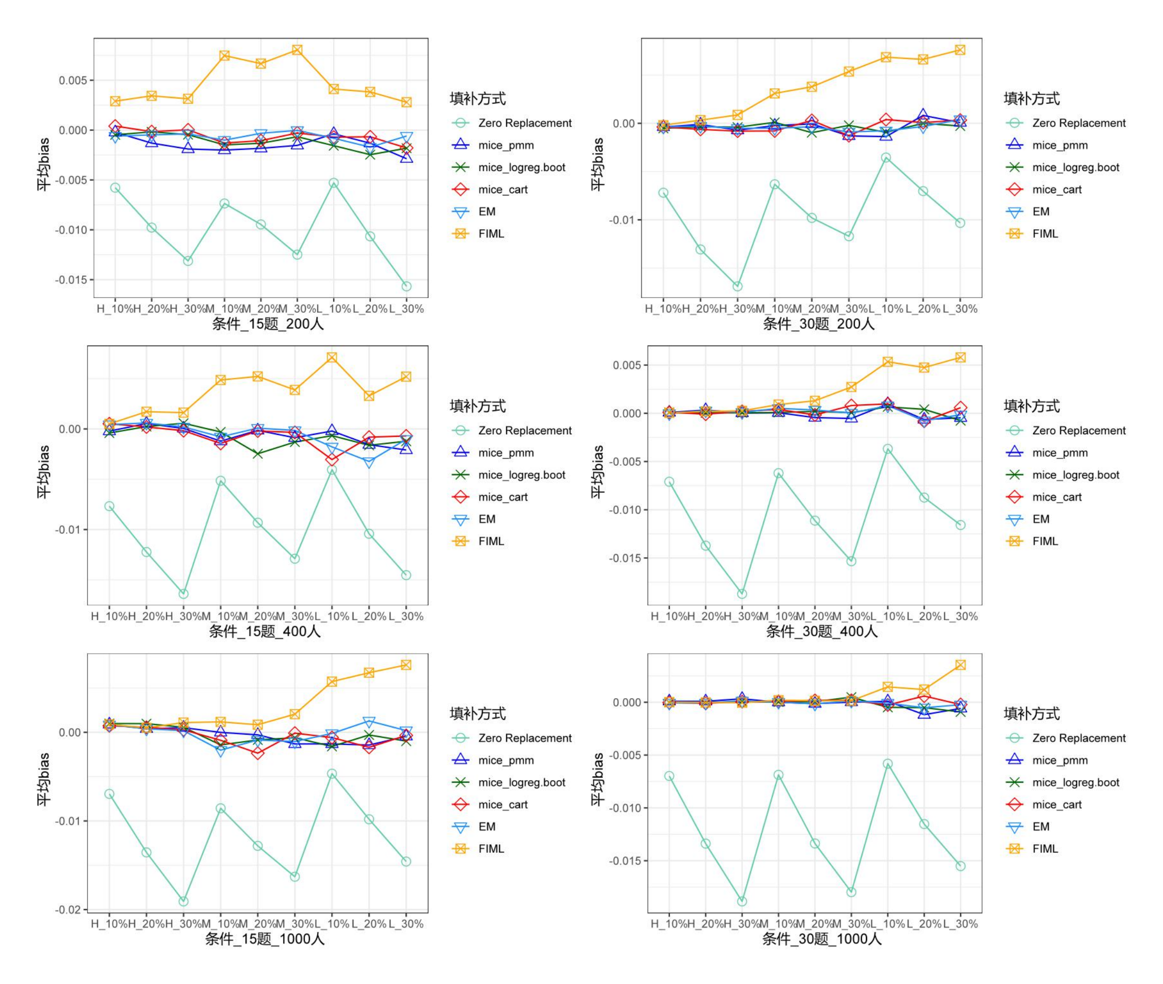


图 1 不同处理方法下题目参数的 Bias（MAR 机制）

注：横坐标条件中第一个字母表示题目质量（H:高质量，M：中等质量，L：低质量），第二个数字表示缺失率(10%，20%，30%)。Zero Replacement 代表零替换法，mice-pmm, mice-logreg.boot, mice-cart 依次代表了多重插补中的预测均值匹配，基于 Bootstrap 的贝叶斯逻辑回归和分类回归树法，EM 代表期望最大化法，FIML 代表了全息极大似然估计法。

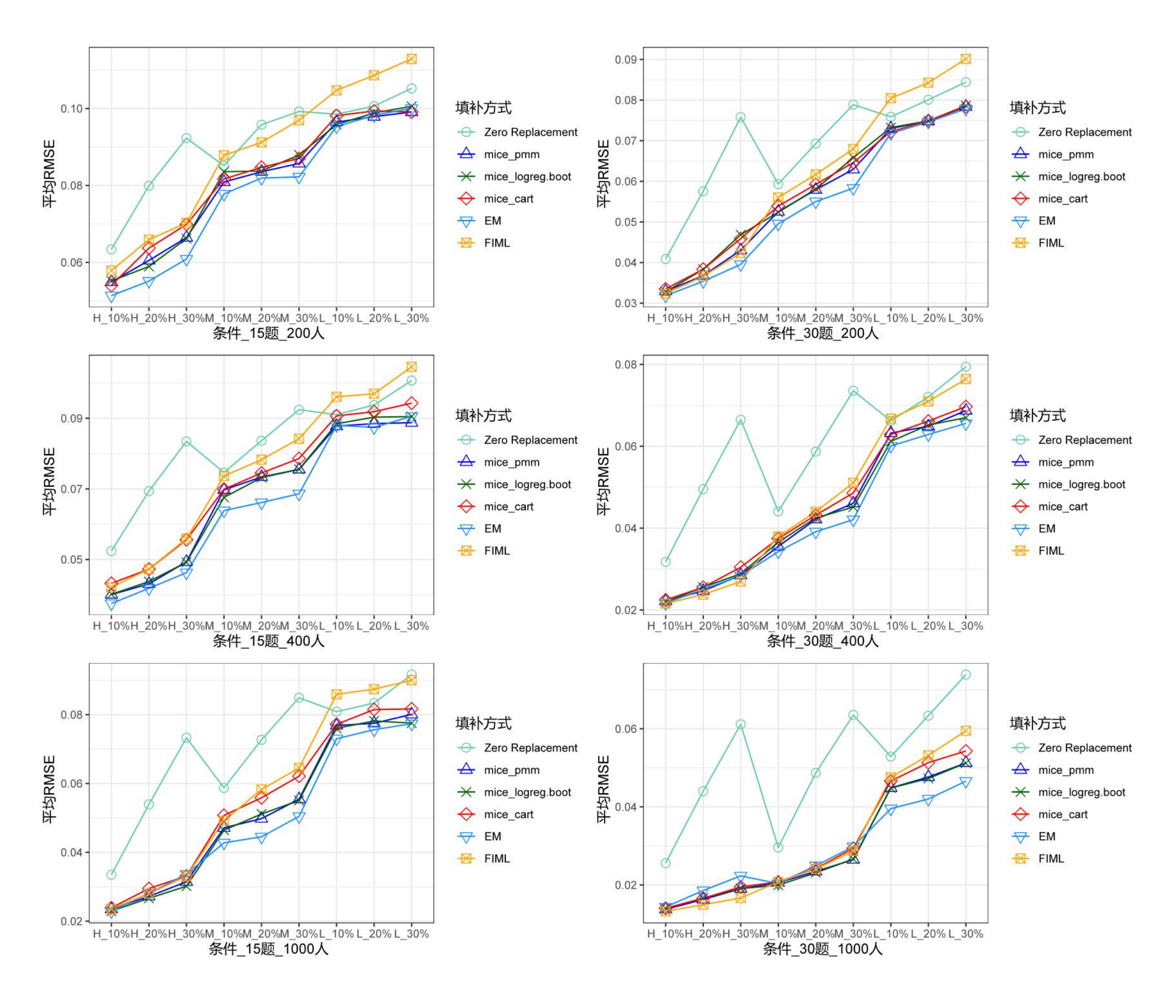


图 2 不同处理方法下题目参数的 RMSE（MAR 机制）

（2）PCCR 估计结果与讨论

图 3 呈现了 MAR 机制下模式判准率的估计结果，共包含 324 个条件。随着被试和题目数量的增多，题目质量的提高和缺失率的降低，各方法的模式判准率均会增大。

总体而言，EM 和 FIML 表现最好，其后依次为 MI 和 ZR。首先，EM 法的 PCCR 在多数条件下最高，其范围为 0.144~0.855。FIML 和 EM 相似，其 PCCR 的范围为 0.123~0.866。值得注意的是，FIML 估计 PCCR 时表现较好，尽管它在估计题目参数时不是表现最好的方法，例如在 MAR 机制下，FIML 与表现最佳的 EM 法之间的 Bias 差值最大仅为 0.008，RMSE 差值最大仅为 0.014，尤其在题目数量较大、题目质量较高时，其表现尤佳。且 FIML 法的操作比 EM 和 MI 法更加便捷，前者属于“一步式”操作，即无需填充未作答数据，仅用已作答数据即可进行参数估计；后两者属于“两步式”操作，需要先将缺失数据插补出来，之后再使用诊断模型进行参数估计。因此，出于便捷性考虑，可将 FIML 作为首选方法。其次，MI 法表现居中，且 MI 系列中的三种方法的 PCCR 曲线相似度也较高，其 PCCR 的范围为 0.114~0.838，略低于 FIML 和 EM 法，但高于 ZR（PCCR 范围为

0.119~0.819）。

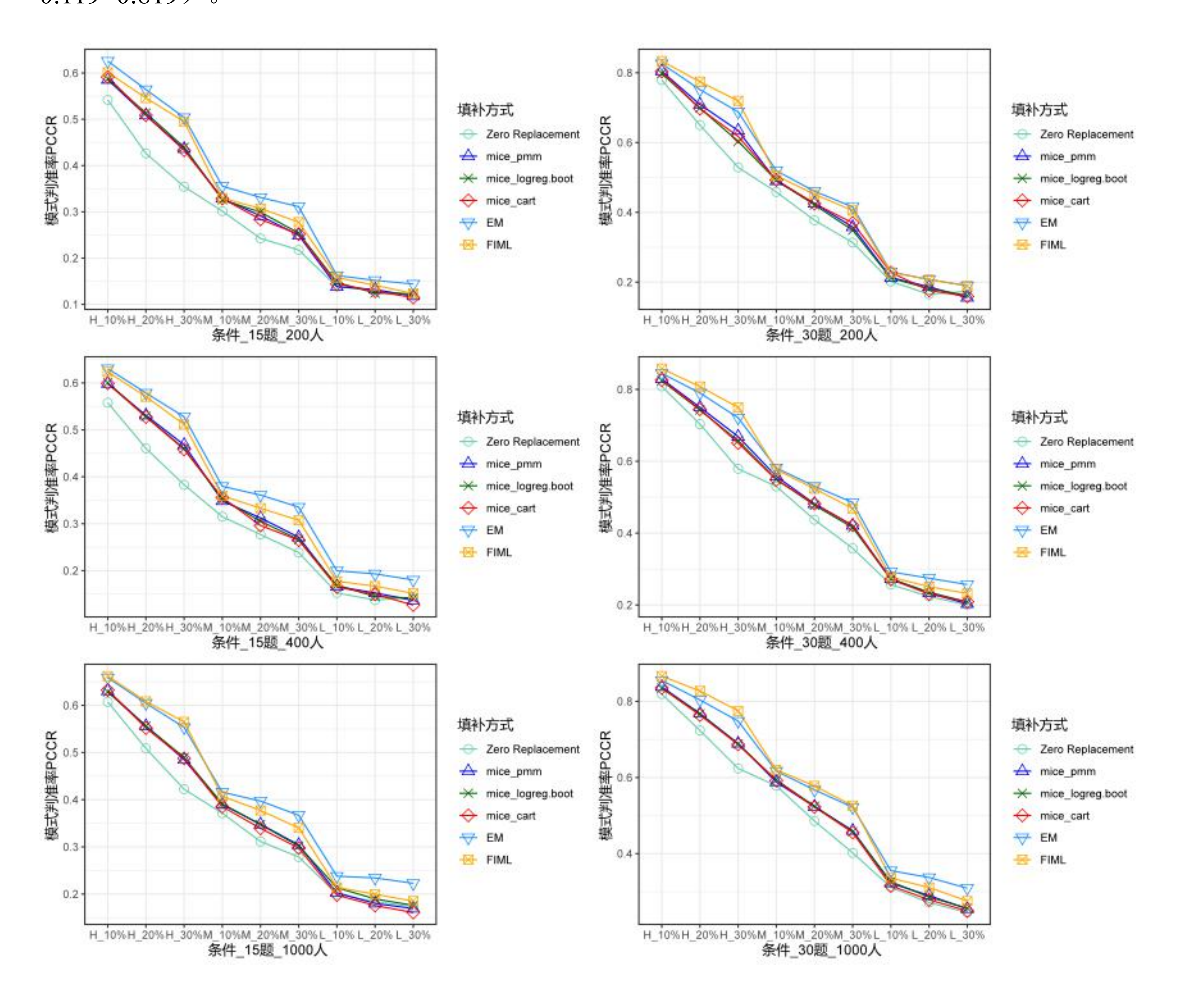


图 3 不同处理方法下题目参数的 PCCR（MAR 机制）

3.4.2 MNAR 机制的结果与讨论

（1）题目参数结果和讨论

图 4 和图 5 呈现了 MNAR 机制不同条件下题目参数的估计结果，共包含 324 个条件。随被试和题目数量的增多，题目质量的提高和缺失率的降低，各方法的题目参数估计精度均在提升。

与 MAR 类似，MNAR 机制下题目参数的 Bias 绝对值和 RMSE 整体较低，表明各方法表现均较好。其中，EM 表现最好，其后依次为 ZR、MI 和 FIML。EM 倾向于高估题目参数，其 Bias 分布范围为－0.003~0.010，RMSE 分布范围为 0.015~0.109。ZR 倾向于低估题目参数，其 Bias 分布范围为－0.011~－0.001，RMSE 分布范围为 0.017~0.099。MI 倾向于高估题目，其 Bias 分布范围为－0.0005~0.010，RMSE 分布范围为 0.015~0.107。FIML 倾向于高估题目参数，其 Bias 值分布范围为 0.002~0.016，RMSE 分布范围为 0.014~0.115。结果显示：首先，与 MAR 机制相比，MNAR 机制下 EM 和 MI 法的 Bias 绝对值增大，由

无偏估计变为高估题目参数，ZR 法的 Bias 绝对值和 RMSE 变小，表现变好。其次，ZR 的表现和 MI 相似但不如 MI 稳定，它主要受题目质量的影响，例如，在 30 题目 400 人、低题目质量、缺失率为 30%时，ZR 法的 RMSE 在所有方法中最小，而相同条件下高题目质量时，ZR 的 RMSE 在所有方法中最大。同 MAR 与 MCAR 机制类似，EM 的总体偏差较小，表现较好，且随着被试量的增多表现变得更好。这也和前文的结果一致，题目质量越高，ZR 表现越差，基于模型的方法表现越好。MI 系列中的三种方法的表现较为相似。

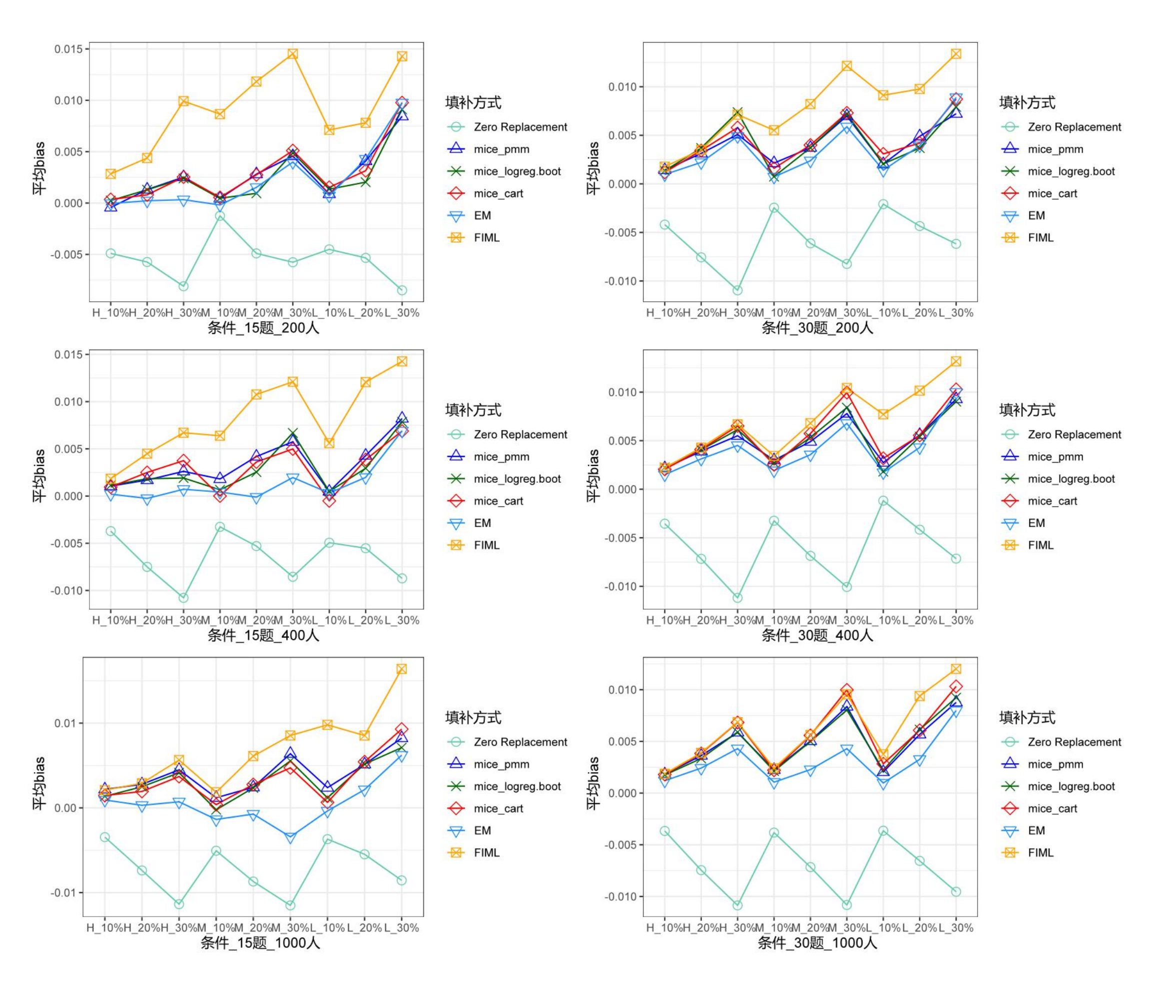


图 4 不同处理方法下题目参数的 Bias（MNAR 机制）

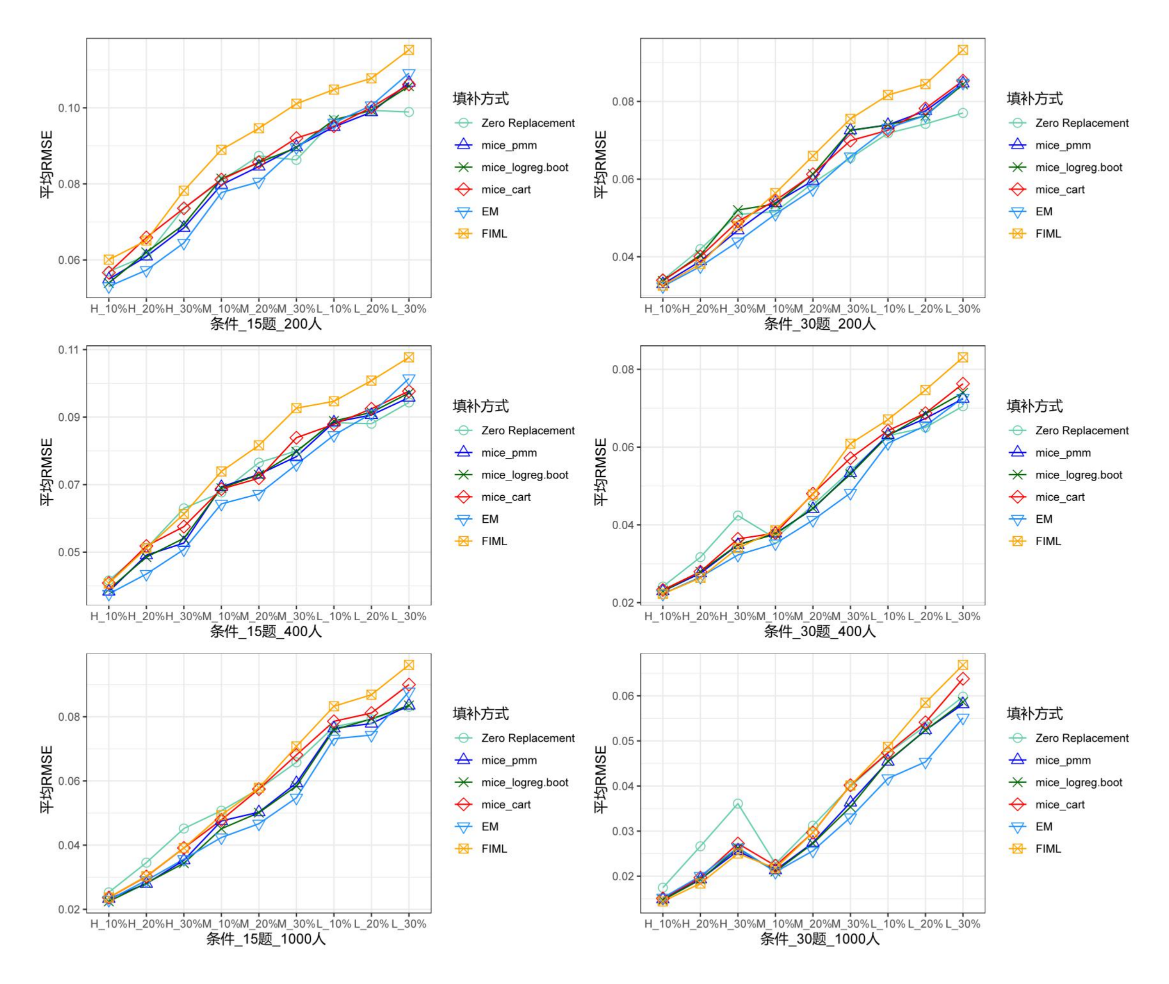


图 5 不同处理方法下题目参数的 RMSE（MNAR 机制）

（2）PCCR 的结果和讨论

图 6 呈现了在 MNAR 机制下题目参数的估计结果，共包含 324 个条件。随被试和题目数量的增多，题目质量的提高和缺失率的降低，各方法的模式判准率均增大。

在 30 题、1000 人、高题目质量且缺失率 30%时，PCCR 最高的 ZR 法与最低的 MI-LOGREG.BOOT 法相差 0.134，因此，与题目参数的估计不同，估计被试 KS 时，各方法间的差异较大。具体而言，EM、FIML 和 ZR 并列为表现最好的方法，MI 次之。其中，EM 法的 PCCR 范围为 0.128~0.857，FIML 的范围为 0.121~0.870，ZR 的范围为 0.111~0.863。MI 和另三种方法相比表现略差但差异不明显，其 PCCR 范围为 0.105~0.843。首先，综合题目参数和 PCCR 的结果，相较于 MAR 和 MCAR 机制，ZR 在 MNAR 机制下表现更好，这一现象的原因可能是：MNAR 机制下，缺失数据对应的原始作答为“0”（即答错）的可能性更高，即认为缺失的产生是由于被试无法作答，与被试的知识掌握状态有关；而 ZR 方法正好使用“0”替换缺失值，同样将缺失看作是由于被试不会作答产生的。因此，使用“0”替换缺失数据的 ZR 法正符合 MNAR 的缺失原理，ZR 法在 MNAR 机制下的表现更好。

其次，与 MAR 机制类似，几乎在所有条件下，MI 的系列方法对被试 KS 和题目参数估计结果均较为相似，这是因为 MI 系列方法均基于 MI 框架进行缺失数据插补。

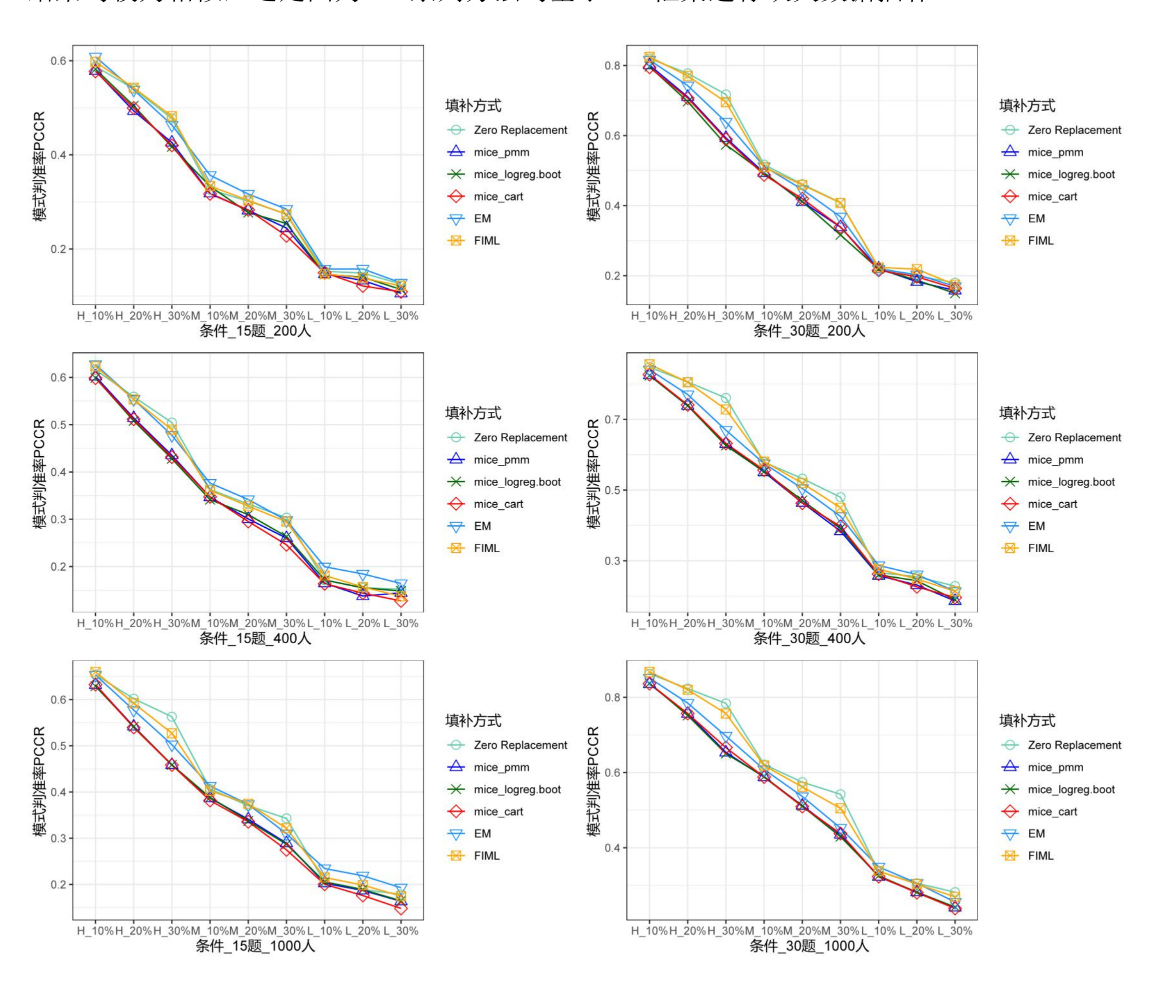


图 6 不同处理方法下题目参数的 PCCR（MNAR 机制）

4 实证研究

4.1 研究数据

为进一步探讨不同缺失值处理方法的生态效度，本研究参考 Shan 和 Wang(2020)的实证研究，使用了 PISA2015 年基于计算机测评的数学测验数据作为实证数据，主要原因为：①缺失比例合适，能够展现出不同缺失值处理方法之间的差异。若缺失率较小，不同缺失值处理方法得到的效果可能差异不会很明显；缺失率较大时（如 30%），所有的缺失值处理方法均表现较差，此时的比较也没有任何意义。②具备已标定好的 Q 矩阵。③属于大型测验，结果可靠。数据包含了九道题目，这些题目在 PISA2015 中的题号分别为 CM033Q01, CM474Q01, CM155Q01, CM155Q04, CM411Q01, CM411Q02, CM803Q01, CM442Q02 和 CM034Q01，题目作答结果均为二分变量。这些题目共考察了四个属性：区别与联系(α_1)、数量(α_2)、空间与形状(α_3)和不定性与数据(α_4)。实证 $\boldsymbol{Q}$ 矩阵参考了 Shan 和

Wang(2020)的研究，参见附录四。本研究选择了多米尼加共和国的 735 名被试进行分析，被试作答结果中，0 表示作答错误，1 表示作答正确，5~9 表示作答缺失。该数据集的缺失比率在各题目上的分布从 0~24.08%不等，总缺失率为 14.02%，缺失比例适中。使用模拟研究中的六种方法对该数据集中的缺失数据进行处理，并采用 GDINA 模型进行估计。

4.2 评价指标

由于实证数据中的 KS 和题目参数真值未知，无法使用模拟研究的评价指标，因此，采用以下几个评价指标：1）相对拟合指标：偏差(Deviance)、赤池信息准则(Akaike information Criterion, AIC; Akaike, 1974)和贝叶斯信息准则(Bayesian information criterion, BIC; Schwarz, 1978)。2）绝对拟合指标：Limited-information statistic M_2 和 root mean square error of approximation ($RMSEA_2$)(Liu et al., 2016)。3）其他指标：题目参数估计标准误(Standard Error, SE)和相关性(Correlation, Cor)。

各指标中，Deviance、AIC 和 BIC 值越小，数据与模型拟合效果越佳，表明经该方法处理的缺失值效果更好。M_2 和 $RMSEA_2$ 都是衡量模型与数据的拟合程度的指标，这两个指标越小，表明模型拟合结果越好。此外，对于 $RMSEA_2$，Liu 等(2016)认为，0.045 是模型良好拟合的标准，0.03 是最佳拟合的标准。SE 指模型估计所得题目参数的标准误，SE 越小，表明题目参数估计结果的离散程度越小，数据越稳定。在估计 SE 时，采用不同的信息矩阵会得到不同精度的结果(Liu et al., 2021)。本研究采用 GDINA 包中的经验交叉相乘方法计算 SE，该方法的优点是操作便捷，且估计参数时表现较好，在 CDM 研究中也较常使用（De la Torre, J, 2009; Nájera et al., 2021; Xu et al., 2020)。相关性指被试在测验上的原始得分与其估计的属性掌握数量之间的相关性，该指标的原理是，被试属性掌握数量越多，其原始得分理应越高(郭磊, 周文杰, 2021)。使用某一种方法对缺失值进行处理后，得到的相关性指标越高，说明缺失值处理效果越好。

4.2 实证研究结果与讨论

实证研究估计得到的各指标结果如表 1（相关性和相对拟合指标）和表 2（绝对拟合指标）所示。就相关性指标而言，EM 的相关性最高，为 0.809，表明这种方法处理缺失数据的效果最佳；其次是 FIML 和 ZR 法，相关性分别为 0.808 和 0.804，但它们的相关性仅略低于 EM，表明它们处理缺失数据的效果基本相似；之后依次为 MI-LOGREG.BOOT、MI-PMM 和 MI-CART，相关性分别为 0.800、0.793 和 0.756。Deviance 分布范围为 4169.45~4633.08，AIC 分布范围为 4235.45~4694.79，BIC 分布范围为 4387.25~4846.59。其中，FIML 的 Deviance、AIC 和 BIC 值均最小，表明拟合效果最好，之后依次是 MI-

LOGREG.BOOT、EM、ZR、MI-CART 和 MI-PMM 法。SE 指标分布范围为 0.243~0.268，其中 MI-PMM 法的 SE 最小，之后依次是 ZR、EM、FIML、MI-LOGREG.BOOT 和 MI-CART 法，表明 MI-PMM 的题目参数估计稳定性表现最好。在绝对拟合指标中，EM、ZR 和 MI-PMM 的 M_2 值的 p 值均大于.05，表明对这批实证数据的拟合效果较好；EM、ZR 和 MI-PMM 的 $RMSEA_2$ 均小于 0.03，表明其拟合效果更佳。

综合各项指标（选取了每项指标上表现最好的三种方法，用“✔”表示，并呈现了各方法得到“✔”的总数和排名），如表 3 所示。在各个指标的表现上，EM、FIML、MI-PMM 三者均在某一个或多个指标上表现最好。从表 3 汇总结果看，EM 在所有指标上表现均较好，是最佳选择。ZR 和 FIML 方法次之，然后为 MI-PMM、LOGREG.BOOT 和 MI-CART，该结果与 NMAR 机制下的实验结果类似。实证研究与模拟研究的 MNAR 机制下得到的结果很相似。PISA 作为大型国际测验，十分受到重视，被试由于个人或环境原因退出测验的可能性较小，由于不会作答而放弃造成作答缺失的可能性较大，这也是 ZR 法表现较好的原因之一。同时这也与 Shan 和 Wang(2020)的研究结果相符。她们运用引入题目层面的缺失数据机制的 CDM 对阿尔巴尼亚共和国的数据进行分析，发现数据的缺失机制更接近 MNAR 机制。此外，MI 系列方法在模拟研究中表现相似，但在实证研究中差异较大，说明选用 MI 系列方法时需要结合实际数据进行模型拟合验证，并根据拟合结果进行选择。

综上，实证研究进一步支持了模拟研究结果，所探讨的缺失数据处理方法具有较高的生态效度。

表 1 实证研究结果 1

处理方法	参考指标				
	Cor	Deviance（-2LL）	AIC	BIC	SE
ZR	0.804	4345.98	4411.98	4563.77	0.256
MI-PMM	0.793	4633.08	4648.78	4800.58	**0.243**
MI-LOGREG.BOOT	0.800	4170.47	4347.45	4499.24	0.263
MI-CART	0.756	4628.56	4694.79	4846.59	0.268
EM	**0.809**	4343.13	4409.13	4560.93	0.258
FIML	0.808	**4169.45**	**4235.45**	**4387.25**	0.260

表 2 实证研究结果 2

处理方法	绝对拟合指标				
	M_2	df	p	$RMSEA_2$	90%CI
ZR	16.69	12	**0.162**	0.023	[0,0.047]
MI-PMM	**13.81**	12	**0.313**	**0.014**	[0,0.042]
MI-LOGREG.BOOT	22.54	12	0.032	0.035	[0.01,0.056]
MI-CART	22.14	12	0.036	0.034	[0.009,0.056]

EM	17.19	12	**0.143**	0.024	[0,0.048]
FIML	22.64	12	0.031	0.035	[0.01,0.057]

表 3 实证研究结果汇总

处理方法	参考指标							✔的总数	排序
	Cor	-2LL	AIC	BIC	SE	*p*	RMESA$_2$		
ZR	✔				✔	✔	✔	4	2
MI-PMM					✔	✔	✔	3	3
MI-LOGREG.BOOT		✔	✔	✔				3	3
MI-CART								0	4
EM	✔	✔	✔	✔	✔	✔	✔	**7**	**1**
FIML	✔	✔	✔	✔				4	2

5 结论与展望

5.1 研究结论

（1）缺失数据会对认知诊断估计产生影响，缺失率的增大会导致所有方法的 PCCR 和题目参数估计精度下降。此外，随着被试与题目数量的减少和题目质量的下降，所有方法的 PCCR 均下降，Bias 绝对值和 RMSE 均上升，表现变差。

（2）整体而言，所有方法都能得到较为精确的题目参数估计值，不同方法间差异不大。其中，在 MAR/MCAR 机制下，EM 的表现最好，其后依次为 MI、FIML 和 ZR 法；在 MNAR 机制下，EM 表现最好，其后依次为 ZR、MI 和 FIML。

（3）估计被试 KS 时，不同方法间 PCCR 差异较大。MAR/MCAR 机制下，EM 和 FIML 表现最好，其后依次为 MI 和 ZR；MNAR 机制下，EM、FIML 和 ZR 并列表现最好，MI 次之。

5.2 方法选择建议

综合模拟与实证研究结果，本研究建议首选 EM 或 FIML 方法。各方法中，EM 在各个指标上均表现较好，是最推荐的方法。但 EM 需要先插补后估计，而 FIML 无需插补便可估计，且 FIML 得到的 PCCR 也较高，尽管 FIML 在估计题目参数时表现不如其它基于模型的方法，但仅是相对而言，其 Bias 绝对值和 RMSE 也均较小，且与同条件下 EM 的表现差异很小。因此，若出于一步到位的处理角度来看，可以优先考虑使用 FIML 进行缺失数据处理。同时，在缺失机制为 MAR 或者 MCAR，以及测验长度较短情况下，研究者应避免使用 ZR 法处理缺失值。

5.3 研究局限及展望

本研究将目前表现效果更好的基于模型的缺失数据处理方法引入 CDA 中，对不同的

缺失数据处理方法进行全面比较，并提出了实践中处理缺失数据的建议。但仍有一些局限，如本研究仅关注了 0-1 计分测验形式，未考虑多级计分情况，而多级计分在现实中也很常见，且能提供更加丰富的作答信息。未来研究可以在多级计分测验中，探究不同缺失数据处理方式对估计结果的影响。其次，近年来纵向 CDA 受到了研究者们的关注(Zhang & Wang, 2018; Kaya & Leita, 2017)，且纵向 CDA 中也存在数据缺失问题，因此，如何处理纵向 CDA 中的缺失数据值得进一步探究。此外，本研究使用了经验交叉相乘法计算实证数据的题目参数标准误。但一些研究指出在估计题目参数标准误时，观察信息矩阵及三明治信息矩阵也是常用且有效的方法（刘彦楼等, 2016; Liu et al., 2019）。因此，在后续研究中可以在缺失值领域进一步对比三种信息矩阵的表现，选取更适合的方法计算标准误。最后，本研究虽然对三种缺失机制分别进行了分析，但实际测验中数据的缺失机制往往不明确，未来研究可以进一步结合包含缺失机制判定的 CDM(Shan & Wang, 2020)，研究实际测验情境下的缺失数据处理模型。

参考文献

Akaike, H. (1974). A new look at the statistical model identification. *IEEE transactions on automatic control, 19(6)*, 716−723.

Aryadoust, V., & Goh, C.C. (2001). Exploring the relative merits of cognitive diagnostic models and confirmatory factor analysis for assessing listening. *Exploring Language Frameworks,* 405.

Bai, S. (2020). Developing a learning progression for probability based on the GDINA model in China. *Frontiers in Psychology, 11*, 2561.

Brown, N. J. S., Svetina, D., & Dai, S. (2014). Impact of methods of scoring omitted responses on achievement gaps. In *annual meeting of the National Council on Measurement in Education, Philadelphia, PA.*

Chen, L., Savalei, V., & Rhemtulla, M. (2020). Two-stage maximum likelihood approach for item-level missing data in regression. *Behavior research methods*, *52*, 2306−2323.

Dai, S. (2017). *Investigation of missing responses in implementation of cognitive diagnostic models* (doctorial dissertation). Indiana University.

De Ayala, R. J., Plake, B. S., & Impara, J. C. (2001). The impact of omitted responses on the accuracy of ability estimation in item response theory. *Journal of educational measurement, 38*(3), 213−234.

De la Torre, J. (2009). DINA model and parameter estimation: A didactic. *Journal of Educational and Behavioral Statistics, 34*, 115−130.

De La Torre, J. (2011). The generalized DINA model framework. *Psychometrika, 76*(2), 179–199.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Methodological), 39*(1), 1–22.

Dong, Y., & Peng, C. Y. J. (2013). Principled missing data methods for researchers. *SpringerPlus, 2*(1), 1–17.

Eekhout, I., Enders, C. K., Twisk, J. W., de Boer, M. R., de Vet, H. C., & Heymans, M. W. (2015). Analyzing incomplete item scores in longitudinal data by including item score information as auxiliary variables. *Structural Equation Modeling: A Multidisciplinary Journal, 22*(4), 588–602.

Enders, C. K. (2010). *Applied missing data analysis*. Guilford press.

Finch, H. (2008). Estimation of item response theory parameters in the presence of missing data. *Journal of Educational Measurement, 45*(3), 225–245.

Gao, X., Wang, D., Cai, Y., & Tu, D. (2018). Comparison of CDM and its selection: A saturated model, a simple model or a mixed method. *Journal of Psychological Science*, *41*(03), 727–734.

[高旭亮, 汪大勋, 蔡艳, 涂冬波. (2018). 认知诊断模型的比较及其应用研究:饱和模型、简化模型还是混合方法. *心理科学*, *41*(03), 727–734.]

Graham, J. W. (2009). Missing data analysis: Making it work in the real world. *Annual review of psychology*, *60*, 549–576.

Graham, J. W., Olchowski, A. E., & Gilreath, T. D. (2007). How many imputations are really needed? Some practical clarififications of multiple imputation theory. *Prevention Science, 8(3),* 206–213.

Guo L., & Zhou W. (2021). Nonparametric methods for cognitive diagnosis to multiple-choice test items. *Acta Psychologica Sinica, 53*(9), 1032–1043.

[郭磊, 周文杰. (2021). 基于选项层面的认知诊断非参数方法. *心理学报, 53*(9), 1032–1043.]

Huisman M., Molenaar I.W. (2001) Imputation of Missing Scale Data with Item Response Models. In: A. Boomsma, M. A. J. van Duijn, T. A. B. Snijders (Eds.), *Essays on Item Response Theory.*

Jeličić, H., Phelps, E., & Lerner, R. M. (2010). Why missing data matter in the longitudinal study of adolescent development: Using the 4-H Study to understand the uses of different missing data methods. *Journal of youth and adolescence, 39*(7), 816–835.

Junker, B. W., & Sijtsma, K. (2001). Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Applied Psychological Measurement, 25(3)*, 258–272.

Kaya, Y., & Leite, W. L. (2017). Assessing change in latent skills across time with longitudinal cognitive diagnosis modeling: An evaluation of model performance. *Educational and psychological measurement, 77*(3), 369–388.

Leacy, F. P., Floyd, S., Yates, T. A., & White, I. R. (2017). Analyses of sensitivity to the missing-at-random assumption using multiple imputation with delta adjustment: application to a tuberculosis/HIV prevalence survey with incomplete HIV-status data. *American journal of epidemiology, 185*(4), 304–315.

Lee, Y. S., Park, Y. S., & Taylan, D. (2011). A cognitive diagnostic modeling of attribute mastery in Massachusetts, Minnesota, and the US national sample using the TIMSS 2007. *International Journal of Testing, 11*(2), 144–177.

Liu, Y. , Tian, W., & Xin, T. . (2016). An application of m2 statistic to evaluate the fit of cognitive diagnostic models. *Journal of Educational and Behavioral Statistics, 41(1),* 3–26.

Liu, Y., Xin, T., & Jiang, Y. (2021, June). Structural parameter standard error estimation method in diagnostic classification models: *Estimation and application.multivariate behavioral research*, 1–20.

Liu, Y., Xin, T., Andersson, B., & Tian, W. (2019). Information matrix estimation procedures for cognitive diagnostic models. *British Journal of Mathematical and Statistical Psychology, 72*, 18–37.

Liu, Y., Xin, T., Li, L., Tian, W., & Liu, X. (2016) Improved method for differential item functioning detection in cognitive diagnosis models: an application of Wald statistic based on observed information matrix. *Acta Psychologica Sinica, 48*, 588–598.

[刘彦楼, 辛涛, 李令青, 田伟, 刘笑笑. (2016). 改进的认知诊断模型项目功能差异检验方法——基于观察信息矩阵的 Wald 统计量. *心理学报, 48*, 588–598.]

Ma, W., & C., & de la Torre, J. (2016). Model similarity, model selection, and attribute classification. *Applied Psychological Measurement, 40(3)*, 200–217.

Ma, W., & de la Torre, J. (2020). GDINA: An R package for cognitive diagnosis modeling. *Journal of Statistical Software*, *93*(14), 1–26.

Ma, W., & de la Torre, J. (2016). A sequential cognitive diagnosis model for polytomous responses. British *Journal of Mathematical and Statistical Psychology, 69*(3), 253–275.

Mazza, G. L. , Enders, C. K. , & Ruehlman, L. S. . (2015). Addressing item-level missing data: a comparison of proration and full information maximum likelihood estimation. *Multivariate Behavioral Research, 50(5),* 504–519.

Marshall, A., Altman, D. G., Royston, P., & Holder, R. L. (2010). Comparison of techniques for handling missing covariate data within prognostic modelling studies: a simulation study. *BMC medical research methodology, 10*(1), 1–16.

Nájera, P., Abad, F. J., & Sorrel, M. A. (2021). Determining the number of attributes in cognitive diagnosis modeling. *Frontiers in psychology, 12*, 321.

Newman, D. A. (2003). Longitudinal modeling with randomly and systematically missing data: A simulation of ad hoc, maximum likelihood, and multiple imputation techniques. *Organizational research methods, 6*(3), 328–362.

Pan, Y., & Zhan, P. (2020). The Impact of Sample Attrition on Longitudinal Learning Diagnosis: A Prolog. *Frontiers in psychology, 11*, 1051.

Rasmussen, K. (2007). *Encyclopedia of measurement and statistics 1*. Sage.

Rezvan, P. H., Lee, K. J., & Simpson, J. A. (2015). The rise of multiple imputation: a review of the reporting and implementation of the method in medical research. *BMC medical research methodology, 15*(1), 1–14.

Rotnitzky, A. (2008). Inverse probability weighted methods. In *Longitudinal data analysis* (pp. 467–490). Chapman and Hall/CRC.

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, *63*(3), 581–592.

Schafer, J. L., & Graham, J. W. (2002). Missing data: our view of the state of the art. *Psychological methods, 7*(2), 147.

Schwarz, G. (1978). Estimating the Dimension of a Model. *The Annals of Statistics, 6(2)*, 461–464.

Shan, N., & Wang, X. (2020). Cognitive diagnosis modeling incorporating item-level missing data mechanism. *Frontiers in psychology, 11*, 3231.

Van Buuren, S. (2018). *Flexible imputation of missing data*. CRC press.

Van Buuren, S., & Groothuis-Oudshoorn, K. (2011). mice: Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software, 45(3)*, 1–67.

Wothke, W. (2000). Longitudinal and multigroup modeling with missing data. *In Modeling longitudinal and multilevel data* (pp. 205–224). Psychology Press.

Xiao, J., & Bulut, O. (2020). Evaluating the performances of missing data handling methods in ability estimation from sparse data. *Educational and Psychological Measurement, 80*(5), 932–954.

Xu, X., & von Davier, M. (2006). Cognitive diagnosis for NAEP proficiency data. *ETS Research Report Series*, *2006*(1), i–25.

Xu, X., de la Torre, J., Zhang, J., Guo, J., & Shi, N. (2020). Estimating CDMs using the slice-within-gibbs sampler. *Frontiers in psychology, 11*, 2260.

Ye, S. J., Tang, W. Q., Zhang, M. Q., & Cao, M. C. (2004). Techniques for missing data in longitudinal studies and its application. *Advances in Psychological Science, 22*(12), 1985−1994.

[叶素静, 唐文清, 张敏强, 曹魏聪. (2014). 追踪研究中缺失数据处理方法及应用现状分析. *心理科学进展*, *22*(12), 1985−1994.]

Zhang, S., & Wang, S. (2018). Modeling learner heterogeneity: A mixture learning model with responses and response times. *Frontiers in psychology, 9*, 2339.

Comparison of missing data handling methods in cognitive diagnosis: zero replacement, multiple imputation and maximum likelihood estimation

SONG Zhilin[1], GUO Lei[1,2], ZHENG Tianpeng[3]

([1] Faculty of Psychology, Southwest University, Chongqing 400715, China)

([2] Southwest University Branch, Collaborative Innovation Center of Assessment toward Basic Education Quality, Chongqing 400715, China)

([3] Collaborative Innovation Center of Assessment toward Basic Education Quality at Beijing Normal University, Beijing 100088, China)

Abstract

The problem of missing data is common in research, and there is no exception for cognitive diagnostic assessment (CDA). Some studies have revealed that both the presence of missing values and the selection of different missing data processing methods would affect the results of CDA. Therefore, it is necessary to attach more attention to the problem in CDA and choose appropriate methods to deal with it. Although the problem in CDA has been explored before, previous studies did not consider multiple imputation (MI) and full information maximum likelihood (FIML), which are widely used in the field of missing data analysis. Moreover, previous studies neglected the comparison using empirical data and saturation models such as GDINA model. In summary, the main purpose of this study are to introduce MI and FIML into CDA, thus making a comprehensive comparison of different missing data handling methods, and further putting forward suggestions for handling missing data in practice.

Simulation study considered six factors: (1) Sample size: 200 participants, 400 participants, and 1000 participants; (2) Test length: 15 test items and 30 test items; (3) Quality of items: high quality, medium quality, and low quality; (4) Missing data mechanisms: missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR); (5) Missing rate: 10%, 20%, and 30%; (6) Missing data handling methods: zero replacement (ZR), MI-CART, MI-PMM, MI-LOGREG.BOOT, Expectation-Maximization algorithm (EM), and FIML. The GDINA model was used, and the analysis process was realized by the GDINA package in R software. Secondly, the PISA 2015 computer-based mathematics data were applied to compare the practical value of the proposed methods.

The results of simulation study revealed that: (1) Missing data results in a decrease in

estimation accuracy. The absolute value of Bias and RMSE both increased and PCCR values of all methods decreased as the sample size, test length and the quality of the items decreased and the missing rate increased; (2) When estimating item parameters, EM performed best, followed by MI. Meanwhile, FIML and ZR methods were unstable; (3) When estimating the KS of participants, EM and FIML performed best as the missing data mechanism was MAR or MCAR. When the missing data mechanism was MNAR, EM, FIML and ZR performed best. The empirical study results further supported the simulation research results. It showed that: (1) For all empirical indicators, EM, FIML, and MI-PMM perform best on one or more indicators; (2) The results obtained under the empirical study and simulation study under the MNAR mechanism are very similar; (3) EM performs well on all indicators, and ZR and FIML methods are slightly worse than EM, followed by MI-PMM, LOGREG.BOOT and MI-CART.

In addition, based on the research results, the following suggestions were provided: (1) EM and FIML should be the first choice. However, if researchers do not want to get the complete data set, FIML could be used as a priority for missing data handling; (2) When the missing data mechanism was MAR or MCAR and the test length was not enough, researchers should avoid using the ZR method to deal with missing data. Finally, this paper ends with the prospects of future researches: (1) The multilevel scoring situation should also be studied; (2) The effectiveness of these methods should be tested in longitudinal research; (3) The performance of more methods of information matrix can be further compared in calculating the standard error to handle missing data; (4) Future research could focus on the missing mechanisms of data onto the real data.

Keywords cognitive diagnosis, GDINA model, missing data, multiple imputation, maximum likelihood estimation

附录一 模拟研究 Q 矩阵

表 1 模拟研究 Q 矩阵（5 属性 15 题目条件）

属性	题目														
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	1	0	0	0	0	1	0	0	0	0	1	1	1	1	0
2	0	1	0	0	0	0	1	0	0	0	1	0	0	0	1
3	0	0	1	0	0	0	0	1	0	0	0	1	0	0	1
4	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
5	0	0	0	0	1	0	0	0	1	1	0	0	1	1	0

表 2 模拟研究 Q 矩阵（5 属性 30 题目条件）

属性	题目														
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	1	0	0	0	0	1	0	0	0	0	1	1	1	1	0
2	0	1	0	0	0	0	1	0	0	0	1	0	0	0	1
3	0	0	1	0	0	0	0	1	0	0	0	1	0	0	1
4	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
5	0	0	0	0	1	0	0	0	1	1	0	0	1	1	0
属性	题目														
	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
1	0	0	0	0	0	1	1	1	1	1	1	0	0	0	0
2	1	1	0	0	0	1	1	1	0	0	0	1	1	1	0
3	0	0	1	1	0	1	0	0	1	1	0	1	1	0	1
4	1	0	1	0	1	0	1	0	1	0	1	1	0	1	1
5	0	1	0	1	1	0	0	1	0	1	1	0	1	1	1

附录二 生成缺失数据的具体步骤

MCAR 机制的数据与其他因素均无关。首先，将该条件下的总缺失率（如 0.1，0.2 或 0.3）设为每个被试在每道题目上的目标缺失率。然后，对每个被试的每一个作答，都会生成一个服从均匀分布 *U*（0,1）的值，并将其与被试的目标缺失率进行比较。若这一数值小于等于目标缺失率，将当前作答替换为缺失，反之则保留原始作答结果。

MAR 机制下缺失数据与已观测到的变量有关，而与产生缺失的变量本身无关。首先，为每个被试生成一个服从标准正态分布 *N*（0,1）的代理变量，这个变量在现实情景中可能是能力、年龄、学习程度等，是对缺失可能性造成影响的个体变量。一般情况下被试的代理变量值越大，在某道题目上的目标缺失率就越低。其次，根据生成的代理变量将被试划分为六个分数段，为每个分数段的被试分配相应的目标缺失率，保证代理变量得分越大，目标缺失率越小。且使所有目标缺失率的平均值等于该条件的总缺失率（即 0.1、0.2 或 0.3）。对每个被试的每一个作答，再生成一个服从均匀分布 *U*（0,1）的值，并将其与被试的目标缺失率进行比较。若其小于目标缺失率，将当前作答替换为缺失，反之不进行处理，保留原始作答结果。

MNAR 机制下缺失数据与缺失前被试是否能正确作答该题目有关，而与其他条件无关。在完整数据基础上，为每个作答分配目标缺失率，原始数据中的错误作答有更高的目标缺失率，正确作答的被试则有更低的目标缺失率。并使所有题目的目标缺失率的均值等于该条件下的总缺失率（即 0.1、0.2 或 0.3）。例如，以一个包含十位被试的数据集为例，如果在完整数据中有五位被试对目标题目作答正确，五位被试作答错误，该条件下缺失率为 0.15，我们向五位作答错误的被试分配 0.10 的缺失率，向五位作答正确的被试分配 0.20 的缺失率。对于每一个被试的每一个作答，都会生成一个服从均匀分布 *U*（0,1）的值，并将其与被试的作答缺失率进行比较。若均值小于缺失率，将当前作答替换为缺失，反之不进行处理，保留原始作答结果。

附录三 五属性条件下 MCAR 机制的结果

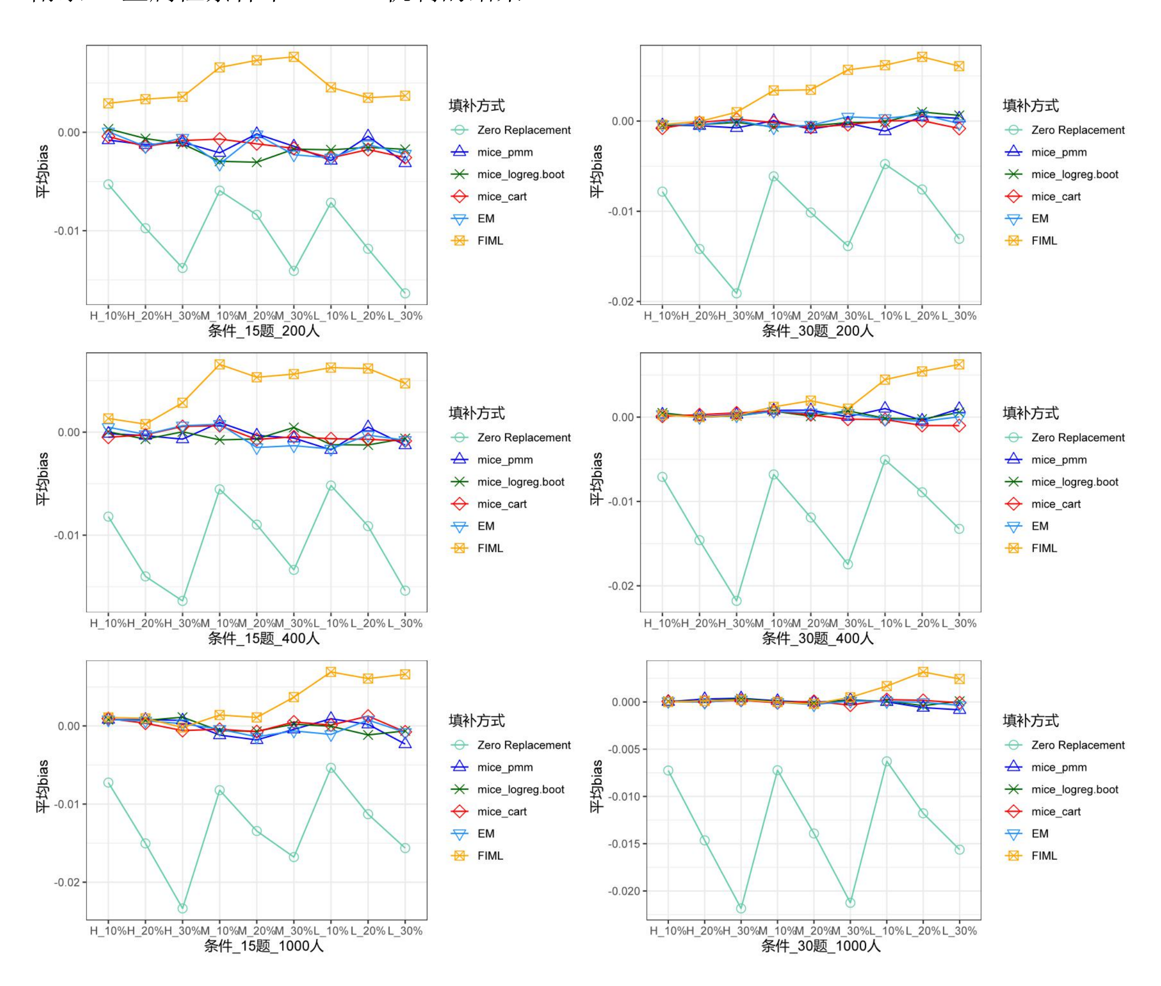


图 1 不同处理方法下题目参数的 Bias（MCAR 机制）

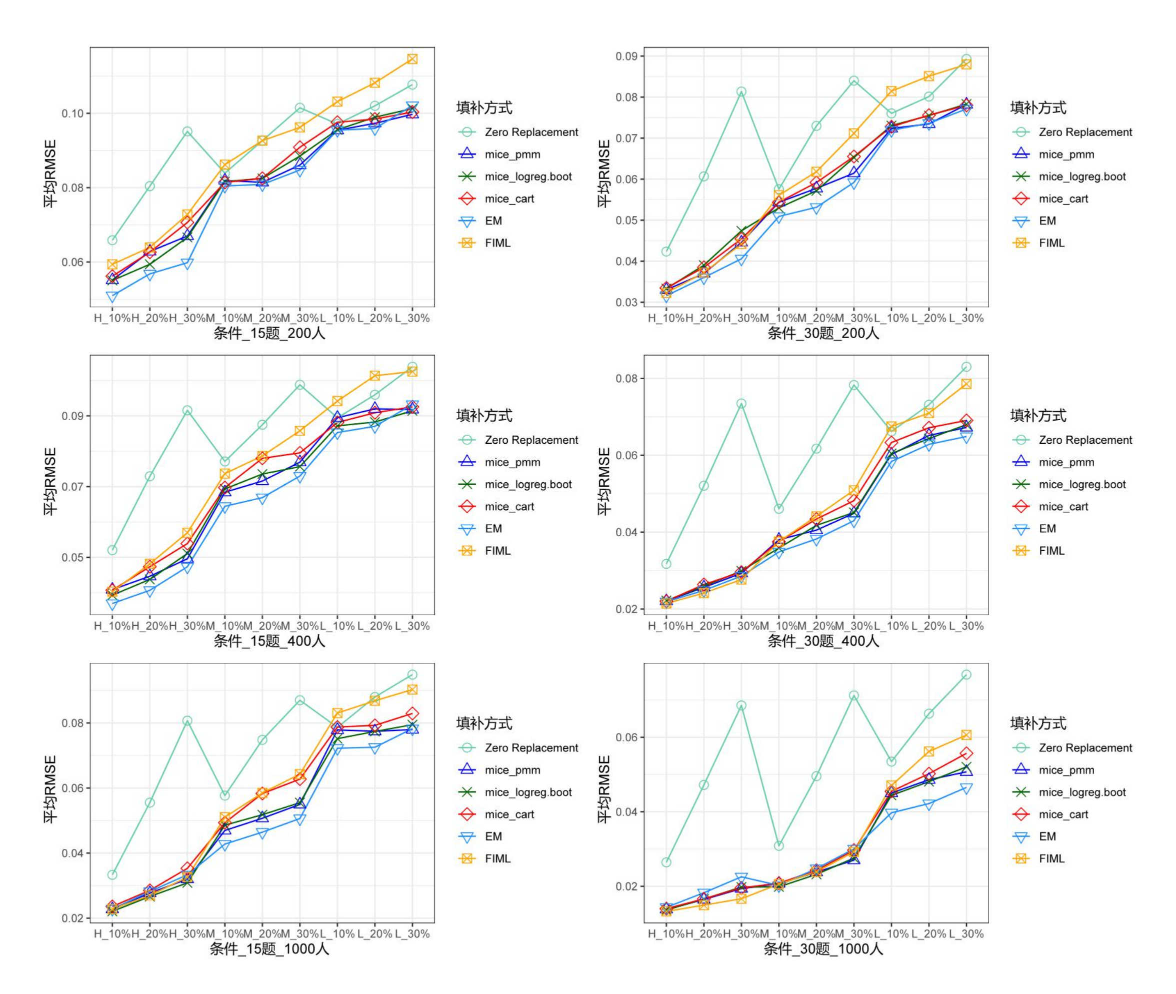


图 2 不同处理方法下题目参数的 RMSE（MCAR 机制）

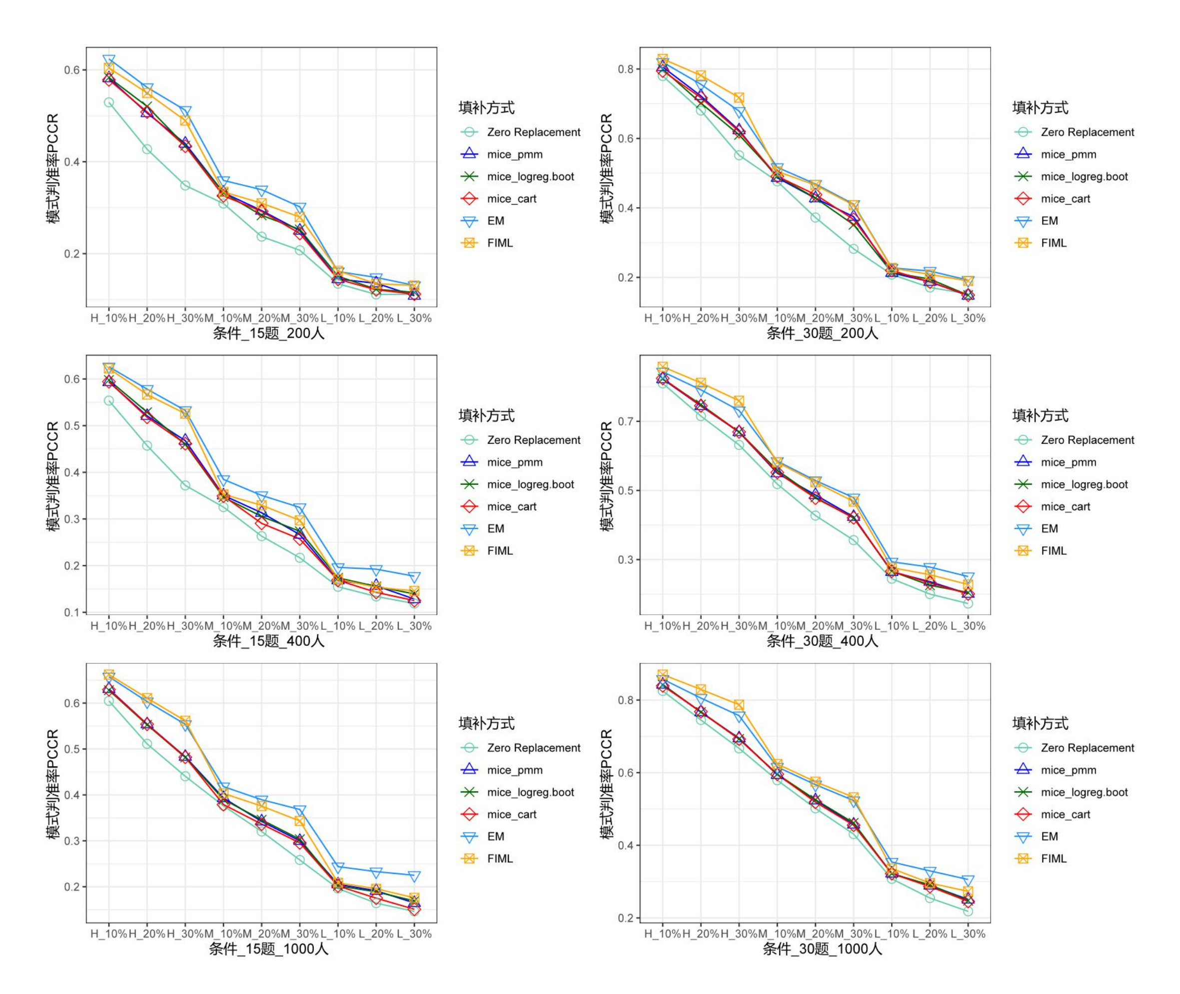


图 3 不同处理方法下题目参数的 PCCR（MCAR 机制）

附录四 实证研究 Q 矩阵

表 3 实证研究 **Q** 矩阵

属性	题目								
	CM033Q01	CM474Q01	CM155Q01	CM155Q04	CM411Q01	CM411Q02	CM803Q01	CM442Q02	CM034Q01
α_1	0	0	1	1	0	0	0	0	0
α_2	1	0	0	0	0	0	0	0	1
α_3	0	1	0	0	1	0	0	1	0
α_4	0	0	0	0	0	1	1	0	0